Users Acting in Mixed Reality Interactive Storytelling

Marc Cavazza[1], Olivier Martin[2], Fred Charles[1], Steven J. Mead[1]
and Xavier Marichal[3]

(1) School of Computing and Mathematics, University of Teesside,
Borough Road, Middlesbrough, TS1 3BA, United Kingdom.
(2) Laboratoire de Télécommunications et Télédétection,
Université catholique de Louvain, 2 place du Levant,
1348 Louvain-la-Neuve, Belgium.
(3) Alterface, 10 Avenue Alexander Fleming, 1348 Louvain-la-Neuve, Belgium.
{m.o.cavazza@tees.ac.uk, martin@tele.ucl.ac.be,
f.charles@tees.ac.uk, xavier.marichal@alterface.com,
steven.j.mead@tees.ac.uk}

"Do you expect me to talk?
Oh no, Mr. Bond. I expect you to die!"
Bond and Auric Goldfinger – from "Goldfinger"

Abstract. Entertainment systems promise to be a significant application for Mixed Reality. Recently, a growing number of Mixed Reality applications have included interaction with synthetic characters and storytelling. However, AI-based Interactive Storytelling techniques have not yet been explored in the context of Mixed Reality. In this paper, we describe a first experiment in the adaptation of an Interactive Storytelling technique to a Mixed Reality system. After a description of the real time image processing techniques that support the creation of a hybrid environment, we introduce the storytelling technique and the specificities of user interaction in the Mixed Reality context. We illustrate these experiments by discussing examples obtained from the system.

1 Rationale

While research in Interactive Storytelling techniques has developed in a spectacular fashion over the past years, there is still no uniform view on the modes of user involvement in an interactive narrative. Two main paradigms have emerged: in the "Holodeck™" approach [10], the user is immersed in a virtual environment acting from within the story; in "Interactive TV" approaches, the user is an active spectator influencing the story from a totally external, "God-mode" perspective [2]. In this paper, we report research investigating yet another paradigm for interactive storytelling, in which the user is immersed in the story but also features as a character in its visual presentation. In this Mixed-Reality Interactive Storytelling approach, the user video image is captured in real time and inserted into a virtual world populated

by autonomous synthetic actors with which he interacts. The user in turn watches the composite world projected on a large screen, following a "magic mirror" metaphor.

In the next sections, we describe the system's architecture and the techniques used in its implementation. After a brief introduction to the example scenario, we discuss the specific modes of interaction and user involvement that are associated with Mixed Reality Interactive Storytelling.

The storytelling scenario supporting our experiments is based on a James Bond adventure, in which the user is actually playing the role of the villain (the "Professor"). James Bond stories have salient narrative properties that make them good candidates for interactive storytelling experiments: for the same reason, they have been used as a supporting example in the foundational work of Roland Barthes in contemporary narratology [1]. Besides, their strong reliance on narrative stereotypes facilitates narrative control and the understanding of the role that the user is allowed to play. The basic storyline represents the early encounter between Bond and the villain (let us call him the Professor). The objective of Bond is to acquire some essential information, which he can find by searching the Professor's office, obtained from the Professor's assistant or even, under certain conditions, (deception or threat) by the Professor himself. The actions of the user (acting as the Professor) are going to interfere with Bond's plan, altering the unfolding of the plot.

The interactive storytelling engine is based on our previous work in character-based interactive storytelling [2]. The narrative drive is provided by the actions of the main virtual character (in this case, the Bond character) that are selected in real-time using a plan-based formalisation of his role in a given scene. The planning technique used is Hierarchical Task Planning, essentially for its representational capabilities [7]. We describe in section 3 how this technique has been adapted to the requirements of Mixed Reality Interactive Storytelling.

2 The Mixed Reality Architecture

Our Mixed Reality system is based on a "magic mirror" paradigm derived from the *Transfiction* approach [4], in which the user's image is captured in real time by a video camera, extracted from his/her background and mixed with a 3D graphic model of a virtual stage including the synthetic characters taking part in the story. The resulting image is projected on a large screen facing the user, who sees his own image embedded in the virtual stage with the synthetic actors (Figure 1).

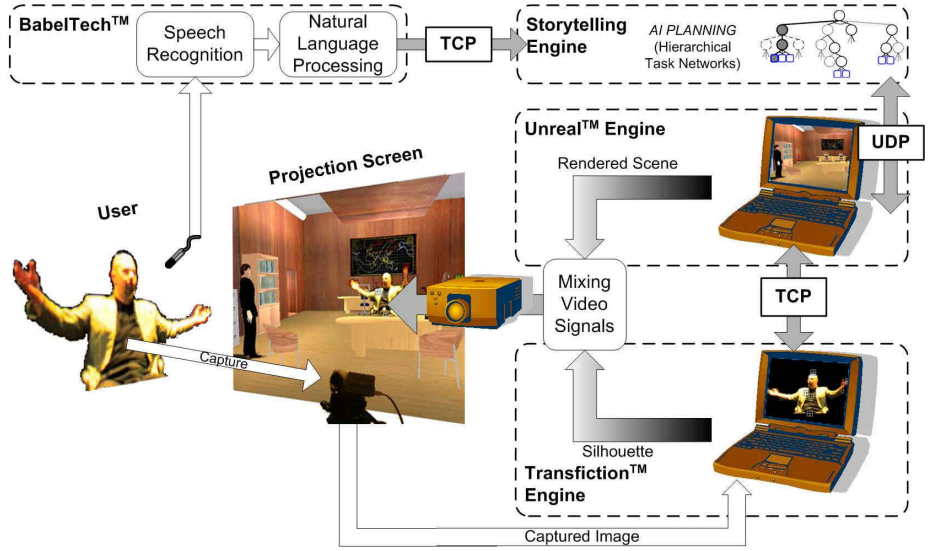

Fig 1. System architecture.

The graphic component of the Mixed Reality world is based on a game engine, Unreal Tournament 2003™. This engine not only performs graphic rendering and character animation but, most importantly, contains a sophisticated development environment to define interactions with objects and characters' behaviours [9]. In addition, it supports the integration of external software, e.g. through socket-based communication.

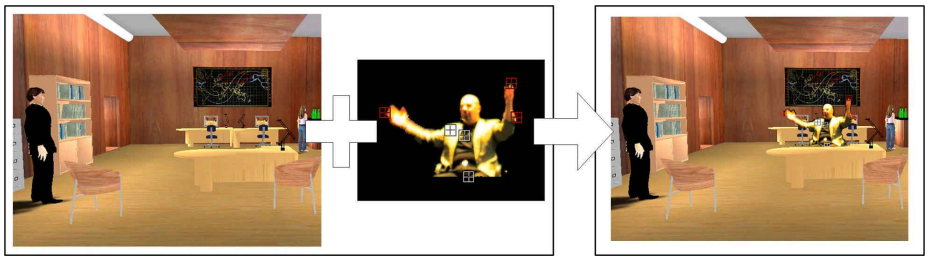

Fig 2. Constructing the Mixed Reality environment.

The mixed environment (Figure 2) is constructed through real-time image processing, using the *Transfiction* engine [6]. A single (monoscopic) 2D camera facing the user analyses his image in real-time by segmenting the user's contours. The objective behind segmentation is twofold: it is intended to extract the image silhouette of the user in order to be able to inject it into the virtual setting on the projection screen (without recurring to chroma-keying). Simultaneously, the extracted body silhouette undergoes some analysis in order to be able to recognise and track the behaviour of the user (position, attitude and gestures) and to influence the interactive narrative accordingly. The video image acquired from the camera is passed to a detection module, which performs segmentation in real time and outputs the

segmented video image of the user together with the recognition of specific points which enable further processing, such as gesture recognition. The present detection module uses a 4×4 Hadamard determinant of the Walsh function and calculates the transform on elements of 4×4 pixels. As a result, it can segment and relatively precisely detect the boundary of objects and also offers some robustness to luminance variations. Figure 3 shows the overview of the change detection process with Walsh-Hadamard transform. First, the module calculates the Walsh-Hadamard transform of the background image. Afterwards, the module compares the values of the Walsh-Hadamard transform of both the current and the background images. When the rate of change is higher than a threshold that has been initially set, this module sets the area as foreground. As segmentation results can be corrupted in presence of shadows (which can be problematic due to variable indoor lighting conditions), we have used invariant techniques [8] to remove such shadows.

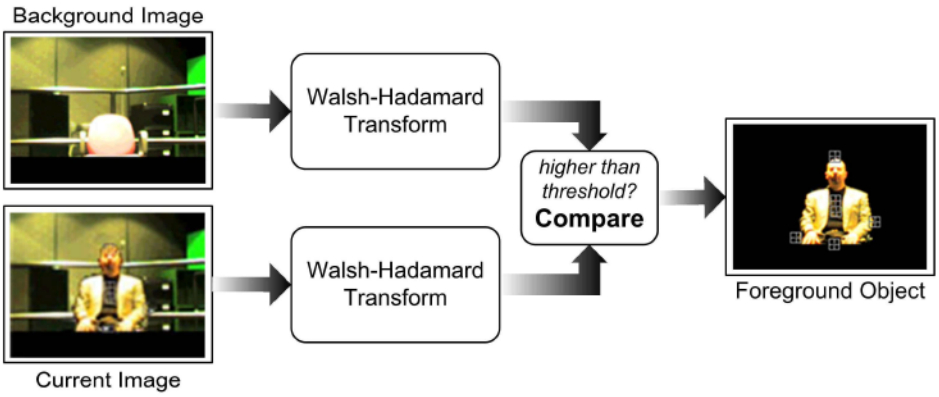

Fig. 3. Extracting the user's image from his background

In this first prototype, the two system components operate by sharing a normalised system of co-ordinates. This is obtained from a calibration stage prior to running the system1. The shared co-ordinates system makes possible to position the user in the virtual image, but most importantly to determine the relations between the real user and the virtual environment. This is achieved by mapping the 2D bounding box produced by the *Transfiction* engine, which defines the contour of the segmented user character, to a 3D bounding cylinder in the Unreal Tournament 2003™ environment, which represents the position of the user in the virtual world (Figure 4) and, relying on the basic mechanisms of the engine, automatically generates low-level graphical events such as collisions and object interaction.

[1] The first prototype does not deal with occlusion in Mixed Reality, which is also set at calibration time. We are currently developing an occlusion management system, which uses depth information provided by the *Transfiction* engine.

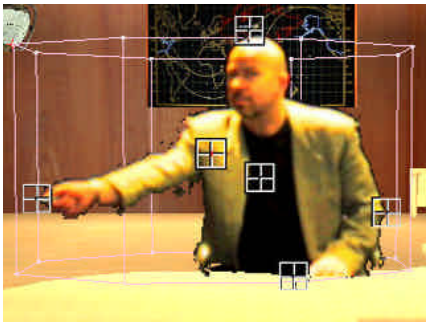

Fig 4. The 3D bounding cylinder determines physical interactions in the Unreal Tournament 2003™ engine.

The two sub-systems communicate via TCP sockets: the image processing module, working on a separate computer sends at regular intervals to the graphic engine two different types of messages, containing updates on the user's position as well as any recognised gestures. The recognised gesture is transmitted as a code for the gesture (plus, when applicable, e.g. for pointing gestures, a 2D vector indicating the direction of pointing). However, the contextual interpretation of the gesture is carried out within the storytelling system.

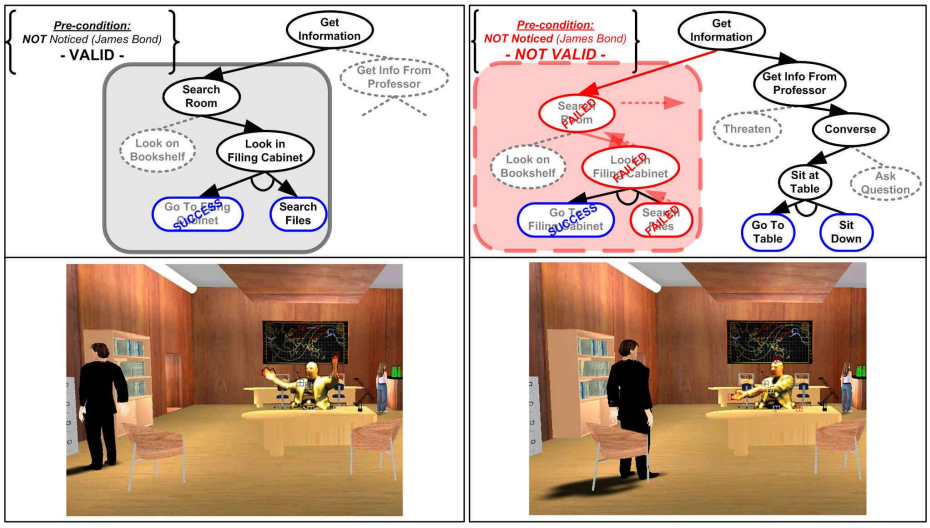

Fig. 5. An Example of User Intervention. The greetings of the user's character force a change of plans in the main character.

To illustrate briefly this implementation of interactive storytelling, we can consider the partial example presented on Figure 5. At this early stage of the plot, Bond has entered the Professor's office and has started searching for documents in the filing cabinet, thinking the room was empty. When the user greets him (with an expressive greeting gesture, as part of his acting), Bond becomes aware of the Professor's

presence and has to direct himself towards him, abandoning his current actions. From that situation, there are many possible instances of his plan (hence the story) depending on the subsequent user's actions, as well as other characters coming into play.

3 User Intervention

In this context, where the user is allocated a role but is left free of his interventions, the specific actions he will take will determine the further evolutions of the plot. In contrast with Holodeck™-like approaches [10], the main character (Bond) is actually a synthetic actor rather than the user, and the storyline is driven by its role. This ensures a spontaneous drive for the story while setting the base for an implicit narrative control. The fact that the user visually takes part in the story presentation obviously affects the modes of user intervention: these will have to take the form of traditional interaction between characters. In other words, the user will have to act. As a consequence, the mechanisms of his normal acting should serve as a basis for interaction.

This fundamental aspects shapes the whole interaction, in particular it determines a specific kind of multi-modal interaction, composed of a spoken utterance and a gesture or body attitude. The latter, being part of the acting actually constitutes a semiotic gesture whose content is complementary but similar in nature to that of the linguistic input [3].

The recognition of a multi-modal speech act comprises an utterance analysed through a body gesture, processed by the *Transfiction* engine described above, and speech recognition. Body gestures from the user are recognised through a rule-based system that identifies gestures from a gesture library, using data from image segmentation that provides in real time the position of user's extremities. One essential aspect of the interaction is that the system is tracking symbolic gestures, which, as part of the user acting, correspond to narrative functions, such as greetings, threatening (or responding to a threat, such as putting his hands up), offering, calling, dismissing, etc.

Fig. 6. Examples of ambiguous gestures.

The gesture recognition process verifies whether first a body gesture has been recognised, then any speech input can provide additional information for the interpretation of the recognised gesture. In our system, speech input is used to help disambiguate gestures, compared to other multimodal approaches, where the gesture is used to disambiguate the speech. Figure 6 illustrates a few potentially ambiguous

body gestures. The correct interpretation of user gestures will be provided by the joint analysis of the user utterance and his gesture.

The speech recognition component is based on the Ear SDK system from BabelTech™, which is an off-the-shelf system including a development environment for developing the lexicon. One advantage is that it can provide a robust recognition of the most relevant topics in context, without imposing constraints on the user (like the use of a specific phraseology) [5]. Finally, after speech and gesture have been combined to produce a multimodal intervention, extra information may be required from the current state of the virtual world, i.e. physical information such as location of objects and characters in relation to the user, etc.

Interactive storytelling has focussed its formalisation efforts on narrative control [11]. It has done so using the representations and theories of narratology. Yet, little has been said about the user's interventions themselves. While they should obviously be captured by the more generic representations of story or plot, there is still a need to devise specific representations for units of intervention.

This formalisation is actually a pre-requisite to successfully map the multi-modal input corresponding to the user acting to the narrative representations driving story generation. In particular, an appropriate mapping should be able to compensate, at least in part, for the limited performance of multi-modal parsing, especially when it comes to speech recognition. The basis for this formalisation is to consider the narrative structure of the terminal actions in the virtual character's HTNs. In previous work [2], we took essentially a planning view to the mapping of user intervention, especially for spoken interaction. This consisted in comparing the semantic content of a user intervention (i.e. a spoken utterance) with the post-conditions of some task-related operator. For instance, if the user provides through spoken interaction the information that a virtual actor is trying to acquire ("the files are on the desk"), this would solve its current goal.

In the current context, we should consider the narrative structure of terminal actions, which formalises explicitly roles for the user and a character. In other words, many terminal actions, such as enquiring about information, have a binary structure with an explicit slot for the user's response. This leads to a redefinition of the character's control strategy in its role application. To account for the fact that user interaction remains optional, all binary nodes (involving a possible user input) should be tested first before attempting a self-contained action from Bond.

One fundamental mechanism by which user actions can be interpreted with a robustness, which exceeds the expected performance of multi-modal parsing, is through the classification of that input using the highest-level categories compatible with interpretation. This approach capitalises on fundamental properties of narrative functions in the specific story genre we are dealing with. If we consider a scene between Bond and the Professor, the majority of narrative functions would develop around a central dimension, which is the agonistic/antagonistic relation.

If we assume that the final product of multi-modal interpretation can be formalised as a speech act, then we can biase the classification of such speech acts towards those high-level semantic dimensions that can be interpreted in narrative terms. The idea is to be able to classify the speech act content in terms of it being agonistic or antagonistic. Each terminal action will in turn have a narrative interpretation in terms

of the user's attitude, which will determine further actions by the virtual Bond character (equivalent to success/failure of a terminal action).

There is indeed a fairly good mapping between speech acts in the narrative context and narrative functions, to the point that they could almost be considered equivalent. Examples of such phenomenon include: denial ("never heard of that, Mr Bond"), defiance ("shoot me and you'll never find out, Mr Bond"), threat ("choose your next witticism carefully …"), etc.

The problem is that this mapping is only apparent at a pragmatic level and, within a purely bottom-up approach, could only be uncovered through a sophisticated linguistic analysis, which is beyond reach of current speech understanding technology. One possible approach is to consider that the set of potential/relevant narrative functions is determined by the active context (i.e., Bond questioning the Professor). And that it is the conjunction of the context and a dimensional feature (i.e. agonistic/antagonistic) that define narrative functions.

For instance, if at any stage Bond is questioning the Professor for information, this very action actually determines a finite set of potential narrative functions: denial, defiance, co-operation, bargaining, etc. Each of these functions can be approximated as the conjunction of the questioning action and a high-level semantic dimension (such as /un-cooperative/, /aggressive/, etc.). The multi-modal analysis can thus be simplified by focussing on the recognition of these semantic dimensions, whose detection be based, as a heuristic, on the identification of surface patterns in the user's utterances, such as ["you'll never"], ["you" … "joking"], ["how would I"].

We illustrate the above aspects within the narrative context where Bond is questioning the Professor in Figure 7. Because of the set of potential narrative functions defined by Bond's current action (i.e. questioning the Professor), a test must be first carried out on the compatibility of the user's action (in this case that the Professor gives away the information) and only after can the character's action be attempted. It should be noted that this does not add any constraints on the user's actions than the one expected, which will impact on the character's plan at another level. In other words, this representation departs from a strict character-based approach to incorporate some form of plot representation, in order to accommodate for the higher level of user involvement.

In the example presented in Figure 7, the joint processing of the gestures and speech leads to interpreting the open arms gesture of the Professor and the identified surface pattern of his utterance *["you" ... "joking"]* as an /un-cooperative/ semantic dimension. Finally, the conjunction of this defined semantic dimension and the current narrative context provide sufficient information to approximate the *denial* narrative function.

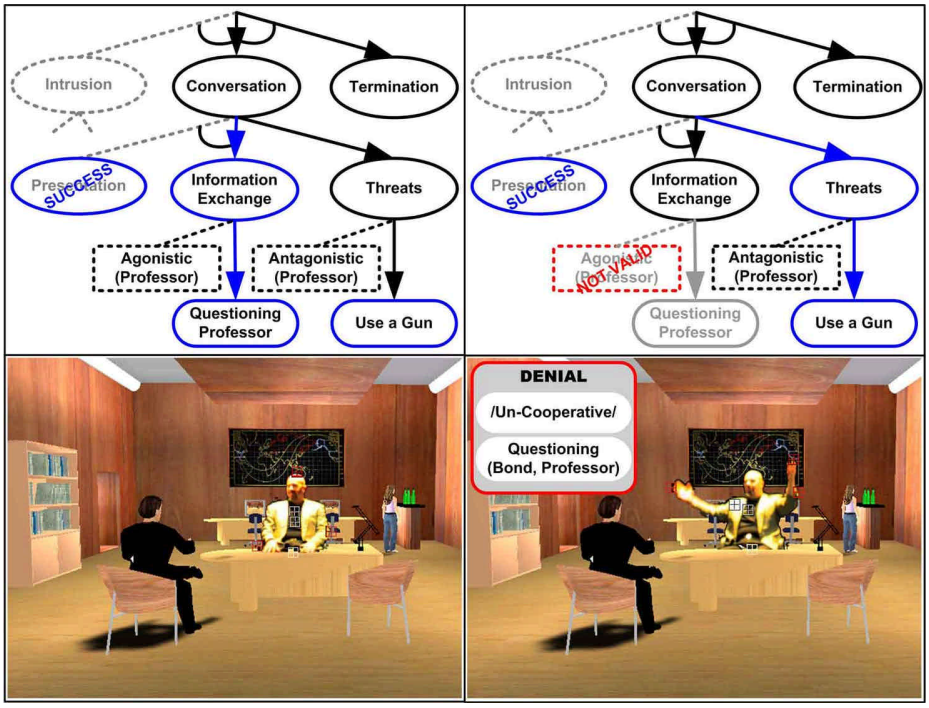

Fig. 7. An Example of Multi-Modal Interaction. Left: Bond is questioning the Professor for information. Right: the Professor replies *"You must be joking, Mr Bond!"* with a corresponding body gesture, denoting a defiance. The multi-modal speech act is interpreted as a *denial*.

4 Conclusions

We have described a first implementation of Mixed Reality Interactive Storytelling, which sets new perspectives on the user's involvement as an actor, which at the same time is also the spectator of the scene within which he is playing. Participating in such narratives potentially only requires that the user is instructed about the baseline story and possible actions, but does not (and should not) require knowledge of Bond's detailed plans and actions, or detailed instructions on his own character's sequence of actions. This work is still at an early stage, and further experiments are mandatory. Our current efforts are dedicated to the integration of robust speech recognition through multi keyword spotting, in order to support natural interaction throughout the narrative.

Acknowledgements

Olivier Martin is funded through a FIRST Europe Fellowship provided by the Walloon Region.

References

1. R. Barthes, Introduction à l'Analyse Structurale des Récits (in French). Communications, 8, pp.1-27, 1966.
2. M. Cavazza, F. Charles, and S.J. Mead, Character-based Interactive Storytelling, IEEE Intelligent Systems, special issue on AI in Interactive Entertainment, pp. 17-24, 2002.
3. M. Cavazza, F. Charles, and S.J. Mead, Under The Influence: Using Natural Language in Interactive Storytelling, 1st International Workshop on Entertainment Computing, IFIP Conference Proceedings, 240, Kluwer, pp. 3-11, 2002.
4. X. Marichal, and T. Umeda, "Real-Time Segmentation of Video Objects for Mixed-Reality Interactive Applications", Proceedings of the "SPIE's Visual Communications and Image Processing" (VCIP 2003) International Conference, Lugano, Switzerland, 2003.
5. S.J. Mead, M. Cavazza, and F. Charles, "Influential Words: Natural Language in Interactive Storytelling", 10th International Conference on Human-Conputer Interaction, Crete, Greece, 2003, Vol 2., pp.741-745.
6. A. Nandi, and X. Marichal, "Senses of Spaces through Transfiction", pp. 439-446 in "Entertainment Computing: Technologies and Applications" (Proceedings of the International Workshop on Entertainment Computing, IWEC 2002).
7. D. Nau, Y. Cao, A. Lotem, and H. Muñoz-Avila, "SHOP: Simple hierarchical ordered planner", Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Stockholm, AAAI Press, 1999, pp. 968-973.
8. E. Salvador, A. Cavallaro, and T. Ebrahimi, "Shadow identification and classification using invariant color models", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001), Salt Lake City (USA), 2001.
9. Special issue on "Game engines in scientific research", Communications of the ACM, 45:1, January 2002.
10. W. Swartout, R. Hill, J. Gratch, W.L. Johnson, C. Kyriakakis, C. LaBore, R. Lindheim, S. Marsella, D. Miraglia, B. Moore, J. Morie, J. Rickel, M. Thiebaux, L. Tuch, R. Whitney, and J. Douglas, "Toward the Holodeck: Integrating Graphics, Sound, Character and Story", in Proceedings of the Autonomous Agents 2001 Conference, 2001.
11. R. Michael Young and Mark Riedl, "Towards an Architecture for Intelligent Control of Narrative in Interactive Virtual Worlds", in Proceedings of the International Conference on Intelligent User Interfaces, January, 2003.